Владо М. Симеуновић
Сања Б. Милић
Универзитет у Источном Сарајеву
Педагошки факултет Бијељина

УДК 159.928.23-057.874
DOI 10.46793/Uzdanica19.1.269S
Оригиналан научни рад
Примљен: 14. јануар 2022.
Прихваћен: 6. мај 2022.

УТВРЂИВАЊЕ САДРЖАЈНЕ ВАЉАНОСТИ НОВОКОНСТРУИСАНИХ ИНСТРУМЕНАТА ИСТРАЖИВАЊА (ПРИМЈЕР ИНСТРУМЕНТА ЗА ПРОЦЈЕНУ ДАРОВИТОСТИ ШКОЛСКЕ ДЈЕЦЕ)[1]

Апстракт: Динамичан научно-технолошки и друштвени развој намеће потребу сталног истраживања његових посљедица у свим сферама живота. Истраживачи широм свијета покушавају да региструју новонастале промјене и на научно верификовани начин утврде њихове ефекте, како би могли предвидјети правце кретања и утицати на будуће догађаје. На том путу се суочавају са низом непознаница које би могле утицати на квалитет резултата који се добијају. Најчешћи проблеми с којима се истраживачи сусрећу су: тешкоће око прецизног одређења предмета истраживања (због мултидисциплинарности проблема), недостатак релевантних података на којима би се спровело истраживање (динамика промјена је велика), те проблеми израде инструмената истраживања (стандардизовани инструменти често нису примјењиви).

Овај рад бави се проблемом одређивања метријских карактеристика новоконструисаних инструмената истраживања у области процјене даровитости дјеце школског узраста који су имплементирани у Софтвер за процјену даровитости. Сматра се да је ваљаност најважнија мјерна карактеристика сваког инструмента. У раду смо показали коришћење садржајне ваљаности као могућег приступа у новоконструисаним инструментима истраживања. Осим теоријских модела чију смо вриједност представили, на практичном примјеру смо показали примјену израчунавања индекса садржајне ваљаности (CVI) и *kappa* статистике, односно приказали смо практични модел примјене постојећих процедура (индекса) за израчунавање садржајне ваљаности и на основу тога закључили да је CVI увјерљив метод за процјену ваљаности садржаја нове (или ревидиране) скале.

Кључне ријечи: садржајна ваљаност, метријске карактеристике, инструменти истраживања, процјена даровитости.

[1] Рад је резултат пројекта „Тестирање софтвера за процјењивање даровитости код ученика” (19.032/961-37/19), који суфинансира Министарство за научнотехнолошки развој, високо образовање и информационо друштво Републике Српске.

vlado.simeunovic@pfb.ues.rs.ba

УВОД

Концепт ваљаности формулисао је Кели (1927: 14), који је изјавио да је тест валидан ако мјери оно за шта тврди да мјери. Ваљаност се често дефинише као мјера у којој инструмент мјери оно што тврди да мјери. У литератури (Крисвол 2005; Палант 2011) познате су методе за утврђивање ваљаности инструмената и то: садржајна, критеријумска и конструкциона анализа.

Стога се процесна операционализација ваљаности састоји од сљедећег:

– пажње истраживача према истраживачком пројекту, који у почетку настоји утврдити домен, дизајн и границе истраживања и усклађеност предмета, дизајна и методе истраживања (тј. критерија за одабир једног или више случајева, подаци који се прикупљају, начини прикупљања и анализа података);

– организације прикупљања података (односно, избор испитаника и информатора, прихватање или одбацивање испитаника, избор докумената итд.);

– кодификације и анализе података која успоставља структурирање концепата, доказа и исправну повезаност категорија;

– анализе података која настоји разумјети феномен кроз више извора података и на потпун начин; и

– дискусије о резултатима и повратка на теорију, чиме се завршава процес теоретизације.

Која од метода ће се примијенити зависи од карактеристика проблема који се истражује. Често се код нових инструмената не може примијенити критеријумска или конструкциона ваљаност, те је нужно да се кроз процес утврђивања садржајне ваљаности учесницима у процесу испитивања језички и логички што више прилагоде поједини ајтеми. Треба осигурати да инструмент укључи одговарајући скуп ставки које утичу на концепт. Што више ставке на скали представљају домен концепта који се мјери, то је већа ваљаност садржаја (Секиран, Бужи 2010).

На примјеру утврђивања садржајне ваљаности приликом израде инструмента за процјену даровитости ученика основних школа показан је квалитет примјене поступка утврђивања садржајне ваљаности. Кориштени су верификовани модели израчунавања садржајне ваљаности (CVR), индекс ваљаности (I-CVR, S-CVI/UA и S-CVI/Ave), као и *kappa* статистика (Флајш, Левин, Чо 1981; Лош 1975; Лин 1986; Полит, Бек 2007; Тилден, Нелсон, Меј 1990).

ВАЉАНОСТ САДРЖАЈА

Ваљаност садржаја први пут је у праксу уведена 1952. године. Ваљаност садржаја се односи на случај у којем се утврђује специфична врста понашања. Обично, тест ће се узорковати из скупа могућих понашања. Академско постигнуће се најчешће испитује за ваљаност садржаја (АПА 1952: 468). Тек је Ленон (1955) суштински увео садржајну ваљаност као легитимни поступак утврђивања ваљаности инструмената истраживања. У овом раду предложено је да се термин 'ваљаност садржаја' користи у смислу како се наводи у АПА стандардима тестирања, наиме да означи обим на који се одговори субјекта могу сматрати репрезентативним узорком његових одговора на ставке теста које представљају стварни или хипотетички скуп ситуација које проистичу из проблема истраживања (Ленон 1955). АПА је у едицији из 1966. године безусловно прихватила садржајну ваљаност као мјеру за валидацију инструмената – „ваљаност садржаја је посебно важна за мјере постигнућа и стручности и за мјере прилагођавања или друштвеног понашања заснованог на посматрању у одабраним ситуацијама" (АПА 1966: 12).

Кромбах се детаљно бавио метриком инструмената и дао потпуно јасно објашњење садржајне ваљаности. Он сматра да валидација садржаја укључује процјену дефиниције домена и процјену колико је тест одговарао тој дефиницији. Пошто се домен често дефинише у смислу теста нацрта, степен до којег је тест у складу са његовим планом описан је као кључни елемент валидације садржаја: „Провјера ваљаности садржаја поставља питање да ли тест одговара нацрту истраживача, и […] да ли би тестни корисник изабрао исти нацрт" (Кромбах 1971: 452).

У међувремену се концепт „садржајне ваљаности" усавршавао, те је Шепард (1993) изнио став да су фундаментални принципи који су у основи ваљаности садржаја опстали. Извршена је модификација терминологије те се умјесто фразе „ваљаност садржаја" уводи „доказ ваљаности у вези са садржајем", који показује степен до којег су узорци ставки, задатака или питања на тесту репрезентативни за неки дефинисани универзум или домен садржаја.

Као што је показано у прегледу литературе, током година постоји консензус да најмање четири елемента квалитета теста дефинишу концепт ваљаности садржаја: дефиниција домена, релевантност домена, репрезентативност домена и одговарајуће процедуре израде теста.

Дефинисање ваљаности садржаја као дефиниција домена, релевантност домена, репрезентативност домена и одговарајуће процедуре конструкције теста илуструју његову разлику од конструктивне ваљаности. Она карактерише његову централну улогу у процјени ваљаности закључака изведених из процеса утврђивања садржаја, а не резултата добијених на тестовима. Ови елементи наглашавају појам да се ваљаност садржаја односи на квалитет

теста. За разлику од конструктивне ваљаности, која се односи на закључке изведене из резултата теста (на тај начин превазилазећи тест), ваљаност садржаја описује потребну компоненту теста. Тестови би требало да буду валидни за садржај. Они би требало да представљају предвиђени домен и не би требало да садрже материјал који је ван тог домена. Дакле, процјена ваљаности садржаја је у великој мјери једнака процјени теста и његових саставних елемената. „Ваљаност" садржаја односи се на кредибилитет, и поузданост самог инструмента за процјену мјерења конструкта од интереса. Ваљаност садржаја је ограниченија у овом смислу од ширег концепта конструктивне ваљаности. Ипак, како можемо да процијенимо закључке засноване на резултату, а да претходно не процијенимо сам инструмент за процјену?

Иако је ваљаност садржаја заснована на тесту, а не на резултату, треба напоменути да она није у потпуности статичан показатељ. Дефиниција домена који се користи за развој теста и карактеристике садржаја теста морају се процијенити у односу на специфичну сврху тестирања. Тест може имати ваљаност садржаја за једну сврху тестирања, али не и за другу. На примјер, садржај теста за процјену даровитости у Босни и Херцеговини не може у потпуности бити прикладан за утврђивање исте у другим крајевима свијета јер се морају узети у обзир језичке и културне претпоставке и карактеристике.

Ваљаност садржаја односи се на способност теста да представља домен задатака, он је дизајниран да мјери исправност дефиниције домена који лежи у основи теста. Дакле, „ваљаност садржаја" је тачан дескриптор пожељних квалитета теста.

Ваљаност садржаја је термин који могу лако разумјети и они који не владају психометријским појмовима. Пречесто се психометричари оптужују да говоре језиком који лаици не разумију. Ипак, свако може да разумије концепт да ли тест адекватно мјери садржај за који је дизајниран. Дакле, ваљаност садржаја је користан термин и требало би да се задржи у ријечнику корисника тестова, теоретичара мјерења и других истраживача.

Фокус је на утврђивању да ли ставке узорковане за укључивање у инструмент адекватно представљају домен садржаја којим се инструмент бави и релевантност домена садржаја за предложену интерпретацију резултата добијених када се мјера користи. Због тога је ваљаност садржаја у великој мјери функција начина на који је инструмент развијен. Када је домен адекватно дефинисан, циљеви који представљају тај домен су јасно објашњени, конструише се исцрпан скуп ставки за мјерење сваког циља, а затим се примјењује поступак случајног узорковања да се изабере подскуп ставки из ове веће групе за укључивање у инструмент, вероватноћа да ће инструмент имати адекватан садржај је велика.

УТВРЂИВАЊЕ САДРЖАЈНЕ ВАЉАНОСТИ

Постоје десетине метода за израчунавање степена до којег су два или више оцјењивача конзистентна или конгруентна у својим оцјенама (Стемлер 2004). Како је Стемлер описао, различите методе се могу класификовати у једну од три категорије: процјене конзистентности, процјене консензуса и процјене мјерења. Предложене алтернативе „садржајне ваљаности инструмента – CVI" су сви индекси који дају конзистентност или консензусне процјене.

Процјене конзистентности се фокусирају на степен до којег су стручњаци доследни (поуздани) у примјени скале оцјењивања за релевантност предмета. Један приступ квантификацији конзистентности је коефицијент алфа, приступ који су предложили Волц и сарадници (Волц, Стрикланд, Ленц 2005) као метод израчунавања ваљаности садржаја када постоји више стручњака. Израчунавање алфа коефицијента узима у обзир све варијације у одговорима стручњака, док CVI смањује варијације сажимањем четири категорије рејтинга у дихотомију релевантно/нерелевантно. Постоје, међутим, проблеми са коришћењем коефицијента алфа (или других математички сличних индекса конзистентности као што је коефицијент корелације унутар класе) у сврхе ваљаности садржаја. Као прво, *алфа* израчунава за све ставке и стручњаке, и на тај начин пружа ограничене информације за оцјењивање појединачних ајтема или појединачних процјењивача. Друго ограничење је то што се висока вриједност *алфа* може добити чак и у ситуацијама у којима је сагласност о ваљаности садржаја ниска. Као што су наведени истраживачи примијетили, висока алфа вредност „не значи да су сви стручњаци додијелили исту оцјену, већ да се релативни редослед или рангирање резултата које је додијелио један стручњак поклапа са релативним редослиједом који су додијелили други стручњаци" (Волц, Стрикланд, Ленц 2005: 156).

Као екстремни примјер, разматрано је шест стручњака који оцјењују четири ставке на скали релевантности 1–4. Претпоставимо да су три експерта дала оцјене 1, 1, 2 и 2 за ставке, а остала три стручњака дала су оцене 3, 3, 4 и 4 истим ставкама. У овој ситуацији, вриједност *алфа* би била 1,00, упркос значајном неслагању међу стручњацима и оцјенама ниске релевантности свих ставки од стране три стручњака. За управо описане оцјене, S-CVI/A не би био 0,50, што би већина сматрала неприхватљиво ниским. Индекси доследности, као што је коефицијент алфа, фокусирају се на интерну доследност (тј. поузданост међу оцјењивачима), а не на сагласност међу стручњацима (Полит, Бек, Овен 2007: 461).

CVI и већина других индекса који су предложени за ваљаност садржаја спадају у категорију коју је Стемлер (2004) назвао процјенама консензуса. Консензусне процјене се односе на степен до којег експерти „дијеле заједничку интерпретацију конструкта" (Стемлер 2004: 2) и могу да се договоре о томе како примијенити скалу оцјењивања на ставке. Низак ниво слагања

вјероватно одражава неуспјех да се демонстрира заједничко разумијевање конструкције.

Широко коришћен приступ за израчунавање консензусне процјене је израчунавање пропорције у сагласности, као у CVI. Критичари CVI-а (и једноставне пропорције у сагласности) највероватније ће бити забринути због могућности превисоких вриједности које узрокују случајна подударања.

Kappa статистика, консензусни индекс слагања међу оцјењивачима који се прилагођава случајном договору, предложен је као мјера ваљаности садржаја. На пример, Вајнд, Шмит и Шефер (2003) користили су и CVI и *kappa* коефицијент са више оцјењивања у валидацији садржаја свог алата за процјену ризика од остеопорозе. Они су тврдили да је *kапа* статистика била важан додатак (ако не и замјена за) CVI јер *kапа* пружа информације о природи степена слагања.

Неколико истраживача предложило је друге методе консензуса за процјену споразума међу оцјењивачима у сврху ваљаности садржаја, као што су резимирали Линдел и Брант (1999). Ово укључује индексе који се називају T (Тинсли, Вајс 1975); rWG (Џејмс, Димари, Волф 1993) и r WG (Линдел, Брант, Витнеј 1999).

Још један индекс је понудио Лош (1975), који је предложио индекс слагања међу оцјењивачима за ставке на скали које се називају однос ваљаности садржаја (CVR), који се користи за дихотомне оцјене ставки. Лош је препоручила усредњавање CVR-а да би се добио укупни индекс ваљаности садржаја. Сви ови индекси врше прилагођавања за случајну сагласност, а сви осим једног (T индекс) пружају дијагностичке информације о обје ставке и укупној скали.

ПРИМЈЕР ИЗРАЧУНАВАЊА САДРЖАЈНЕ ВАЉАНОСТИ КОД ИНСТРУМЕНАТА ЗА ПРОЦЈЕНУ ДАРОВИТОСТИ КОД УЧЕНИКА У ОСНОВНОЈ ШКОЛИ

Разноврсност термина који означавају сложени феномен даровитости указује на одсуство једне униформно прихваћене одредбе, било од стране истраживача, било од стране оних који непосредно раде са даровитима. У педагошкој литератури помиње се стотињак дефиниција даровитости, чија ширина обухвата различита схватања, тако да не постоји консензус о томе како прецизно дефинисати и мјерити даровитост. Избор метода и поступака за идентификацију даровитости свакако зависи од тога на који начин се приступа и тумачи даровитост, али је веома важно да постоји тијесна веза између схватања даровитости, обиљежја даровитих, њихове идентификације и наставних програма намјењених таквој популацији (Џорџ 2005). Истраживачи који се баве феноменом даровитости праве разлику између традиционалне

дефиниције даровитости и модерне, вишедимензионалне, која карактерише даровитост путем когнитивних, квалитативних, психолошких и друштвених аспеката. С обзиром на то да и један и други приступ дефинисању даровитости имају низ предности, препоручује се да се у циљу оптимизације самог поступка идентификације даровитих комбинују оба (Перлет, Зиглер 1997; Хелер и др. 2005).

Оправданост израде инструмената за процјењивање даровитости који су интегрисани у софтвер лежи у драматичним промјенама које су наступиле у друштвено-економском развоју, а које су покренуле размишљање о кључној улози школе и образовања, са нагласком на значај идентификације и развоја даровитих, односно појединаца који су покретачи друштвених промјена и творци најпрогресивнијих идеја које резултирају новим, оригиналнијим и савременијим продуктима људске културе.

Коришћење Гарднерове теорије о мултиплој интелигенцији (1993, 1996) на просторима Босне и Херцеговине оправдано је из више углова. Прва оправданост лежи у самој структури наставног плана и програма у оквиру кога су и наставни предмети подијељени према областима интелигенција које препоручује Гарднер, те су се све четири групе процјењивача ослањале на успјехе у одређеним предметима. Друга оправданост је у томе што ће наставници користећи инструменте за процјену једноставније формирати секције и одабрати ученике за разна такмичења која се организују на овим просторима, а која су исто тако диференцирана према различитим врстама интелигенције. Ослањајући се на теорију мултипле интелигенције наставницима и ученицима ће се отворити могућност да прилагоде стратегије учења најјачим снагама ученика (Милић, Симеуновић 2020).

Процес развоја инструмената одвијао се кроз четири фазе. Прва фаза започета је прегледом релевантне литературе о вишеструким интелигенцијама (Гарднер 1993; Армстронг 2009) и квалитативном анализом садржаја и већ постојећих инструмената намјењених идентификацији даровитости у различитим доменима. У овој фази организован је детаљни полуструктурирани интервју са петнаест родитеља чија деца похађају ниже разреде основних школа, петнаест учитеља и петнаест ученика. Резултати интервјуа водили су идентификовању домена садржаја релевантих за процјену мултипле интелигенције. На почетку је генерисано 270 ајтема који су својим садржајем обухватали свих девет врста мултипле интелигенције (МИ). Коначним прегледом у оквиру истраживачке групе поједини ајтеми су елиминисани или прекомпоновани јер су се понављали или нису били довољно разумљиви. Коначна верзија документа садржавала је 189 ајтема. У другом кораку и након одабира петнаест стручњака за садржај, укључујући стручњаке за развој инструмената (четири особе), стручњаке за истраживање даровитости (шест особа) и пет школских психолога, формирано је стручно вијеће које је донијело квантитативне и квалитативне процјене о сваком ајтему у инструменту.

Од чланова комисије три пута се тражило да одлуче о ваљаности садржаја, индексу ваљаности и свеобухватности инструмента. У сваком обраћању, путем имејла или лично, припремили смо писмо које је садржавало циљеве истраживања са упутством за бодовање. Од стручњака се тражило да одреде да ли је неки ајтем валидан за инструмент. У ту сврху, оцијенили су сваки ајтем од 1 (није валидан), 2 (корисно, али није неопходно) и 3 (валидан). Резултате смо провјерили помоћу формуле CVR = (Ne - N/2)/(N/2) (Лош 1975). У табели коју је представио наведени аутор уз p = .05, прихватају се они ајтеми код којих је CVR = .49 и већи (минимална вриједност CVR за узорак од петнаест стручњака). У овом кругу одбачено је 45 ајтема, тако да је добијена верзија од 144 ставке. У сљедећем кораку израчунат је индекс ваљаности за све појединачне димензије у инструменту (S-CVI/UA и S-CVI/Ave): Verbal-linguistic (S-CVI/UA .77; S-CVI/Ave .89), Logical-mathematical (S-CVI/UA .73; S-CVI/Ave .85), Visual-spatial (S-CVI/UA 0.73; S-CVI/Ave .84), Bodily-kinesthetic (S-CVI/UA .81; S-CVI/Ave .91), Musical (S-CVI/UA .75; S-CVI/ Ave .87), Interpersonal (S-CVI/UA .76; S-CVI/Ave .88), Intrapersonal (S-CVI/UA .73; S-CVI/Ave.84), Naturalistic (S-CVI/UA .74; S-CVI/Ave .86) and Existential (S-CVI/UA .76; S-CVI / Ave .88).

У сљедећем кораку израчуната је садржајна ваљаност (I-CVI) за сваки ајтем у инструменту и модификовани *kappa* индекс [k = (I-CVI-pc) (1-pc)]. У наведеном поступку од укупно 144 ајтема елиминисано је седам чији је I-CVI био испод 0.70 (Лин 1986; Полит, Бек 2007; Тилден и др. 1990), а два ајтема (један из музичко-ритмичке, а други из духовно-филозофске димензије) послати су на ревизију јер им I-CVI био између .70 и .79. Након ревизије коначно су елиминисани. На крају је добијен инструмент са 135 ајтема чији је просјек I-CVI = .84, а модификована *kappa* .72 (Флајш 2003). На основу свега се може закључити да инструмент има задовољавајућу садржајну ваљаност коју би свакако требало провјерити у неком наредном истраживању уз коришћење неке друге методе за утврђивање ваљаности.

Коначан инструмент садржи по 15 ајтема за сваку од девет врста даровитости. Задатак процјењивача је да процијене јачину опажене карактеристике сваком тврдњом на петостепеној скали, која је организована узлазним редоследом од 1 (минимално изражено) до 5 (изразито изражено).

Љествице за процјену вербално-лингвистичке даровитости садрже ајтеме који покривају четири конструкта вербално-лингвистичке интелигенције: вербалну комуникацију, писану комуникацију, способности читања, способности слушања, али и насљедни фактор и мотивацију. Поузданост инструмента провјерена Кронбаховом алфом за сваку групу процјењивача је висока (наставници $\alpha = .881$, родитељи $\alpha = .887$, вршњаци $\alpha = .908$, самопроцјена $\alpha = .913$).

Кронбахов коефицијент за појединачне љествице показао је високу конзистентност на свакој скали:

– инструмент за процјену логичко-математичке даровитости показао је високу конзистентност (наставници α = .884, родитељи α = .873, вршњаци α = .904, самопроцјена α = .910);

– за инструмент за процјену визуелно-просторне даровитости унутрашња конзистентност је висока (наставници α = .876, родитељи α = .876, вршњаци α = .901, самопроцјена α = .916);

– за инструмент за процјену тјелесно-кинестетичке даровитости вриједност Кронбах алфа коефицијента у сва четири случаја (наставници α = .903, родитељи α = .893, вршњаци α = .913, самопроцјена α = .916) јако је висока, што показује изузетно добру поузданост инструмента;

– унутрашња конзистентност инструмента за процјену музичко-ритмичке даровитости је висока за све четири групе (наставници α = .885, родитељи α = .885, вршњаци α = .907, самопроцјена α = .922);

– унутрашња конзистентност инструмента за процјену интерперсоналне даровитости је висока за све четири групе (наставници α = .873, родитељи α = .880, вршњаци α = .904, самопроцјена α = .910);

– за инструмент за процјену интраперсоналне даровитости унутрашња конзистентност инструмента је висока за све четири групе (наставници α = .882, родитељи α = .869, вршњаци α = .907, самопроцјена α = .911);

– унутрашња конзистентност инструмента за процјену природњачке даровитости је висока (наставници α = .882, родитељи α = .869, вршњаци α = .907, самопроцјена α = .911);

– за инструмент за процјену духовно-филозофске даровитости унутрашња конзистентност је висока (наставници α = .873, родитељи α = .863, вршњаци α = .912, самопроцјена α = .906) (Симеуновић, Милић 2020).

ЗАКЉУЧАК

Дискусија о ваљаности представљена у овом раду фокусирала се на питања везана за ваљаност садржаја инструмента. Формулације јединствене концептуализације ваљаности су усредсређене на логичку формулацију и питања везана за закључке који се изводе. Валидација треба да буде усредсређена на јединствену концептуализацију захтјева који се постављају конструкторима инструмената да иду даље од демонстрирања ваљаности теста за одређену сврху. Приказали смо практични модел примјене постојећих процедура (индекса) за израчунавање садржајне ваљаности и на основу тога се може закључити да је CVI увјерљив метод за процјену ваљаности садржаја нове (или ревидиране) скале. Међутим, програмери скале треба да препознају да када користе мјеру међуоцењивања, као што је нпр. CVI, процјењују се сви аспекти ситуације. Ако је вриједност CVI ниска, то може значити да ставке нису биле добре операционализације основног конструкта, да су конструкт

спецификације или упутства стручњацима били неадекватни, или да су сами стручњаци били пристрасни, непосвећени или недовољно компетентни. То подразумијева да на почетку процеса валидације програмери скале морају напорно да раде на развоју квалитених питања и спецификација конструката и да изаберу добар панел стручњака. Као Лин (1986, 1995) и Хајнес и сарадници (1995), ми подржавамо концепт вишеструких итерација у напорима утврђивања високе ваљаности садржаја, што подразумијева ригорозне анализе домена и процеса развоја инструмента.

Прва итерација валидације садржаја би идеално укључивала преглед од стране великог панела стручњака – можда њих 8–15. Експерте треба пажљиво бирати, користећи добро дефинисане критеријуме попут оних које су предложили Грант и Дејвис (1997).

Фокус прве етапе би био на ставкама – откривању које од њих треба ревидирати или одбацити, добијању савјета о томе да ли су потребне додатне ставке да би се адекватно обухватио домен интересовања и улагање напора да се утврди да ли су аспекти конструкције представљени ставкама у исправним пропорцијама. I-CVI добијене вриједности би тада представљале критеријум одлучивања, упућивале на ревизију или одбијања ајтема, а коментари стручњака би водили ка развоју било које нове ставке. Сви резултати I-CVI нешто нижи од .80 сматрали би се кандидатима за ревизију, а они са веома ниским вриједностима кандидатима за брисање.

Осим ако су потребне само мање ревизије ставки на основу резултата првог круга, треба спровести други круг стручног прегледа. У другом кругу, мања група стручњака (нпр. 3–5) може радити на процјени релевантности ревидираног скупа ајтема и на израчунавању S-CVI. Лин (1986) је примјетила да оцјењивачи могу бити извучени из истог круга стручњака као у првом кругу, или могу бити и нови панел. Коришћење подскупа стручњака из првог круга има изразите предности, јер се тада информације из првог круга могу користити за одабир најспособнијих панелиста. На примјер, подаци из првог круга могли би да се користе за елиминисање стручњака који су били досљедно попустљиви (нпр. који су дали оцјене 4 свим ставкама) или досљедно оштри, или чије оцјене нису биле у складу са оцјенама већине других стручњака. Квалитативна повратна информација од стручњака у првом кругу у форми корисних коментара од изузетне је важности за наставак рада.

Када се изабере панел другог круга и добију нове оцјене релевантности за ревидирани скуп ставки, онда се може израчунати S-CVI. Прихватљиве вриједности за C-CVI/UA постаје све теже постићи како се број стручњака повећава. Овај приступ игнорише ризик случајних неслагања, а да не помињемо неслучајна неслагања ако је стручњак пристрасан или је погрешно разумио спецификације конструкције. S-CVI/Ave можда је адекватнији индекс јер избјегава ове проблеме, али и зато што инхерентно утеловљује информације о перформансама сваке ставке кроз функцију усредњавања.

Укратко, препоручујемо да инструмент буде састављен од ставки које имају CVR = .49 и већи, I-CVI и S-CVI/AVe од 0,80 или више, те S-CVI/UA 0.70 и више. Ово захтијева снажан концептуални и развојни рад, добре ајтеме, изванредне стручњаке и јасна упутства стручњацима у вези са основним конструкцијама и задатком оцењивања.

Неколико важних закључака за будуће истраживаче.

1. Ваљаност садржаја се односи претежно на фундаментална питања везана за закључке логичке структуре инструмената, а не на закључке изведене из резултата тестирања.

2. Ваљаност садржаја је неопходан, али не и довољан захтјев да бисмо са сигурношћу говорили о ваљаности закључака.

3. Наука у овом тренутку располаже квалитетним инструментаријем за утврђивање ваљаности садржаја као једне од поузданих метријских вриједности у случајевима када се не може утврдити критеријумска ваљаност.

ЛИТЕРАТУРА

АПА (1952): American Psychological Association, Committee on Test Standards. Technical recommendations for psychological tests and diagnostic techniques: preliminary proposal, *American Psychologist*, 7(8), 461–475, https://doi.org/10.1037/h0056631.

АПА (1966): American Psychological Association, *Standards for Educational and Psychological Tests and Manuals*, Washington, DC: American Psychological Association.

Армстронг (2009): T. Armstrong, *Multiple Intelligences in the Classroom* (3rd ed.), Alexandria, VA: Association for Supervision & Curriculum Development.

Вајнд, Шмит, Шефер (2003): C. Wynd, B. Schmidt, M. Schaefer, Two Quantitative Approaches for Estimating Content Validity, *West J Nurs Res*, 25(5), 508–518, https://doi.org/10.1177/0193945903252998.

Волц, Стрикланд, Ленц (2005): C. F. Waltz, O. L. Strickland, E. R. Lenz, *Measurement in nursing and health research* (3rd ed.), New York: Springer.

Гарднер (1993): H. Gardner, *Multiple intelligence, The theory in practice*, New York: Basic Books.

Гарднер (2006): H. Gardner, *Multiple Intelligences: New Horizons*, New York: Basic Books.

Грант, Дејвс (1997): J. S. Grant, L. L. Davis, Selection and use of content experts in instrument development, *Research in Nursing & Health*, 20, 269–274. DOI: 10.1002/(sici)1098-240x(199706)20:3<269::aid-nur9>3.0.co;2-g

Кели (1927): T. L. Kelley, *Interpretation of educational measurements*, New York: Macmillan.

Крисвол (2005): J. W. Creswell, *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (2nd ed.), New Jersey: Pearson Education.

Кронбах (1971): L. J. Cronbach, Test validation, In: R. L. Thorndike (Ed.), *Educational Measurement*, 2nd ed., Washington, DC: American Council on Education, 443–507.

Ленон (1955): R. Lennon, Standardized Testing, *The bulletin of the National Association of Secondary School Principals*, 39(211), 34–40, https://doi.org/10.1177/019263655503921106.

Лин (1986): M. R. Lynn, Determination and quantification of content validity, *Nurs Res*, 35(6) 382–385, https://doi.org/10.1097/00006199-198611000-00017.

Лин (1995). M. R. Lynn, Development and Testing of the Nursing Role Model Competence Scale (NRMCS), *Journal of Nursing Measurement*, 3(2), 93–108, DOI: 10.1891/1061-3749.3.2.93.

Линдел, Брант (1999): M. K., Lindell, C. J. Brandt, Assessing interrater agreement on the job relevance of a test: A comparison of CVI, T, rWG(J)}, and r*WG(J)} indexes, *Journal of Applied Psychology*, 84(4), 640–647, https://doi.org/10.1037/0021-9010.84.4.640.

Линдел, Брант, Витни (1999): M. K. Lindell, C. J. Brandt, D. J. Whitney, A revised index of interrater agreement for multiitem ratings of a single target, *Applied Psychological Measurement*, 23, 127–135, https://doi.org/10.1177/01466219922031257.

Лош (1975): C. H. Lawshe, A quantitative approach to content validity, *Personnel Psychology*, 28, 563–575, https://doi.org/10.1111/j.1744-6570.1975.tb01393.x.

Милић, Симеуновић (2020): S. Milic, V. Simeunovic, Concordance between giftedness assessments by teachers, parents, peers and the self-assessment using multiple intelligences, *High Ability Studies*, 1–14. DOI: 10.1080/13598139.2020.1832445

Палант (2011): J. Pallant, *SPSS survival manual. A Step by step guide data to analysis using SPSS-4th edition*, Maryborough: Midland Typesetters.

Перлет, Зиглер (1997): C. Perleth, A. Ziegler, Uberlegungen zur Begabungsdiagnose und Begabtenforderung in der Berufsaus- und Weiterbildung [Considerations on diagnosis and promotion of gifted in personnel training and vocational education], In: U. Kittler, H. Metz-Gockel (Eds.), *Padagogische Psychologie in Erziehung und Unterricht*, 100–112.

Полит, Бек (2006): D. F. Polit, C. T. Beck, The content validity index: Are you sure you know what's being reported? Critique and recommendations, *Research in Nursing & Health*, 29, 489–497. DOI: 10.1002/nur.20147

Полит, Бек, Овен (2007): D. F. Polit, C. T. Beck, S. Owen, Is the CVI an Acceptable Indicator of Content Validity? Appraisal and Recommendations, *Research in Nursing & Health*, 30, 459–467. DOI: 10.1002/nur.20199

Секиран, Бужи (2010): U. Sekaran, R. Bougie, *Research methods for business: A skill building approach* (5th ed.), West Sussex, UK: John Wiley & Sons Ltd.

Стемлер (2004): S. Stemler, A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability, *Practical Assessment, Research, and Evaluation*, 9, article 4, https://doi.org/10.7275/96jp-xz07

Тилден, Нелсон, Мej (1990): V. P. Tilden, C. A. Nelson, B. A. May, Use of qualitative methods to enhance content validity, *Nursing Research*, 39(3), 172–175, https://doi.org/10.1097/00006199-199005000-00015.

Тинсли, Вајс (1975): H. E. Tinsley, D. J. Weiss, Interrater reliability and agreement of subjective judgments, *Journal of Counseling Psychology*, 22(4), 358–376, https://doi.org/10.1037/h0076640.

Флајш, Левин, Чо Пејк (2003): J L. Fleiss, B. Levin, M. Cho Paik, *Statistical Methods for Rates and Proportions* (3th ed.), Copyright John Wiley & Sons, Inc.

Хајнес, Ричард, Кубани (1995): S. Haynes, D. Richard, E. Kubany, Content validity in psychological assessment: A functional approach to concepts and methods, *Psychological Assessment*, 7, 238–247, https://doi.org/10.1037/1040-3590.7.3.238.

Хелер, Перлет, Лим (2005): K. Heller, C. Perleth, T. Lim, The Munich Model of Giftedness Designed to Identify and Promote Gifted Students, In: R. Sternberg, J. Davidson (Eds.), *Conceptions of Giftedness*, Cambridge: Cambridge University Press, 147–170.

Џејмс, Димари, Волф (1993): L. R. James, R. G. Demaree, G. Wolf, An assessment of within-group interrater agreement, *Journal of Applied Psychology*, 78(2), 306–309, https://doi.org/10.1037/0021-9010.78.2.306.

Џорџ (2005): D. George, *Obrazovanje darovitih: Kako identificirati i obrazovati darovite i talentirane učenike*, Zagreb: Educa.

Шепард (1993): L. A. Shepard, Evaluating test validity, *Review of Research in Education*, 19, 405–450, https://doi.org/10.3102/0091732X019001405.

Vlado M. Simeunović

Sanja B. Milić

University of East Sarajevo
Faculty of Education in Bijeljina

DETERMINATION OF THE CONTENT VALIDITY OF NEWLY DESIGNED RESEARCH INSTRUMENTS (EXAMPLE INSTRUMENTS FOR ASSESSING THE GIFTEDNESS OF SCHOOL CHILDREN)

Summary: Dynamic scientific-technological and social development imposes the need for constant research of its consequences in all spheres of life. Researchers around the world are trying to register new changes and determine their effects in a scientifically verified way, so that they can predict the directions of movement and influence future events. Along the way, they face a series of unknowns that could affect the quality of the results obtained. The most common problems encountered by researchers are: difficulties in accurately determining the research subject (due to multidisciplinary problems), lack of relevant data on which to conduct research (dynamics of change is considerable), and problems in developing research instruments (standardized instruments are often not applicable).

This paper deals with the problem of determining the metric characteristics of newly constructed research instruments in the field of giftedness assessment of school-age children that are implemented in the Giftedness Assessment Software. Validity is considered to be the most important measurement characteristic of any instrument. The paper reports on the use of content validity as a possible approach in newly designed research instruments. In addition to the theoretical models, the application of content validity index (CVI) and kappa statistics on a practical example are presented, i.e. a practical model of application of existing procedures (index) for content validity calculation. It is concluded that CVI is a reliable method of assessing the validity of the content of a new (or revised) scale.

Keywords: content validity, metric characteristics, research instruments, assessment of giftedness.